FEBS 28792 FEBS Letters 575 (2004) 109–111

Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides

Hedvig Tordai, László Patthy*

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Karolina út 29, P.O. Box 7, Budapest H-1518, Hungary

Received 28 June 2004; revised 19 August 2004; accepted 24 August 2004

Available online 11 September 2004

Edited by Takashi Gojobori

Abstract Analysis of the exon–intron structures of 2208 human genes has revealed that there is a statistically highly significant excess of phase 1 introns in the vicinity of the signal peptide cleavage sites. It is suggested that amino acid sequences surrounding signal peptide cleavage sites are significantly enriched in phase 1 proto-splice sites and this has favored insertion of spliceosomal introns in these sites.

Keywords: Exon–intron structure; Intron insertion; Proto-splice site; Signal peptide

1. Introduction

One of the major enigmas of eukaryotic genome evolution is the proliferation of spliceosomal introns of protein-coding genes. Whereas genes of unicellular eukaryotes contain very few introns, most genes of multicellular eukaryotes contain several introns. In particular, in the vertebrate lineage protein-coding genes usually have numerous introns and the size of intronic DNA substantially exceeds the size of exonic DNA.

Although the mechanisms underlying the proliferation of spliceosomal introns of eukaryotic protein-coding genes are poorly understood, there is now general consensus that spliceosomal introns arose only in the eukaryotic lineage [1–3]. None of the completely sequenced prokaryotic genomes shows any sign of ancient spliceosomal introns, supporting the notion that spliceosomal introns were never present in prokaryotes but arose soon after the origin of the first eukaryotes. It is noteworthy in this respect that all eukaryotic lineages that have been examined so far contain spliceosomal introns and spliceosomal components necessary for their removal [4–7].

There is now strong evidence that in all eukaryotic lineages the loss and gain of spliceosomal introns is an ongoing process [8–10]. To explore the evolutionary dynamics of introns, Rogozin et al. [11] have recently compared the intron positions of 684 orthologous gene sets from eight complete genomes of animals, plants, fungi, and protists and constructed parsimonious scenarios for the evolution of the exon–intron structure of these genes. The inferred evolutionary scenario showed that although the common ancestor of these eukaryotes had numerous introns, most of these ancestral introns have been differentially lost in fungi, nematodes, arthropods and Plasmodium. On the other hand, numerous novel introns have been inserted into vertebrate and plant genes, whereas in other lineages, intron gain was much less prominent.

Although the processes leading to intron gain are still not fully understood, the most plausible, most widely accepted hypothesis is that intron insertion may occur via a process similar to group II intron retrotransposition [12–15]. According to this hypothesis, the spliceosomal components remain transiently associated with a recently excised intron and then attach at a potential splice site of a non-homologous pre-mRNA, where they catalyze the reverse reaction. The modified pre-mRNA is reverse transcribed, the resulting cDNA participates in a recombination with its parent gene thereby inserting a novel intron into the gene. An attractive feature of this mechanism is that it ensures that the inserted sequence has the full complement of intron signature sequences required for efficient splicing. Furthermore, this pathway would guide intron insertion to sites – proto-splice sites – that possess the appropriate exonic features that facilitate efficient splicing [16,17]. Selection thus favors insertion of perfect spliceosomal introns into such proto-splice sites: the inserted intron will be accepted only if it is efficiently spliced out. Insertion in a region that does not conform to the proto-splice-site consensus will not be tolerated.

The recent work by Coghlan and Wolfe [18] provides strong evidence for insertion of novel introns in proto-splice sites. These authors have compared the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* and identified >6000 introns that are unique to one or the other species. Through phylogenetic analysis, they have identified 122 cases where an intron content difference between *C. elegans* and *C. briggsae* was caused by intron insertion rather than deletion. Significantly, novel introns were found to have a much stronger exon splice site consensus sequence (AG/G) than the general population of introns, suggesting that novel introns tend to be inserted at AG↓G, where ↓ is the insertion site. These findings support the conclusion of Qiu et al. [19] that recently gained introns of eukaryotic protein families are inserted into AG↓G sites.

In the present work, we provide evidence that amino acid sequences surrounding signal peptide cleavage sites are significantly enriched in phase 1 proto-splice sites and this has favored insertion of spliceosomal introns in these sites.

* Corresponding author. Fax: +361-4665-465.
E-mail address: patthy@enzim.hu (L. Patthy).

0014-5793/$22.00
doi:10.1016/j.febslet.2004.08.045

2. Results and discussion

Full-length human protein sequences were selected from the Swiss-Prot 41.11 database, http://www.expasy.org/sprot [20], identified by gene name and duplicates were removed. In the resulting database, 1873 sequences with unique gene names were found to possess an N-terminal signal peptide, 6325 sequences lacked a signal peptide. The boundaries of their signal peptide domains were defined using the annotations of the Swiss-Prot database entries.

The human subset of the Exon-Intron Database EID132, http://www.mcb.harvard.edu/gilbert/eid [21], was searched with these sequences and the matching entries were compiled together with information on the position and phase class of their introns. The resulting databases contained 711 genes of signal peptide containing proteins and 1497 genes of proteins lacking a signal peptide (Table 1).

To compare the frequency distribution of phase 0, phase 1 and phase 2 introns along the sequences of human proteins containing or lacking signal peptides, the first 100 residues of both groups were analyzed. The distribution of introns located in the N-terminal 100 residues of human proteins lacking a signal peptide has shown a rather uniform pattern along the sequence (Fig. 1A) for all three intron phase classes. The relative frequency of phase 0 introns (44.6%), phase 1 introns (33.8%) and phase 2 introns (21.6%) in this N-terminal region is quite similar to that observed for the full-length proteins (phase 0 intron: 47.4%, phase 1 intron: 29.9%, phase 2 intron: 22.7%; cf. Table 1). These observations are in harmony with the previous findings that, in general, phase 0 introns are the most abundant (approximately 50%), followed by phase 1 introns (about 30%) with phase 2 introns being the least abundant (about 20%) [22,23].

Table 1
Distribution of introns in the translated regions of human genes encoding proteins that lack or possess N-terminal secretory signal peptides

	Proteins lacking signal peptide	Proteins with signal peptide
Number of protein sequences	1497	711
Total number of introns	11473	4978
Phase 0 intron	5444, 47.4%	1765, 35.5%
Phase 1 intron	3428, 29.9%	2336, 46.9%
Phase 2 intron	2601, 22.7%	877, 17.6%
Introns in the first 100 residues	2483	1210
Phase 0 intron	1107, 44.6%	379, 31.3%
Phase 1 intron	840, 33.8%	604, 49.9%
Phase 2 intron	536, 21.6%	227, 18.8%

The distribution of introns located in the N-terminal 100 residues of human proteins possessing a signal peptide is strikingly different from that observed in the case of proteins lacking a signal peptide (cf. Fig. 1A and B). First, phase 1 introns are significantly more abundant (49.9%) than phase 0 introns (31.36%) or phase 2 introns (18.8%). Second, the distribution of introns along the sequence is far from uniform:

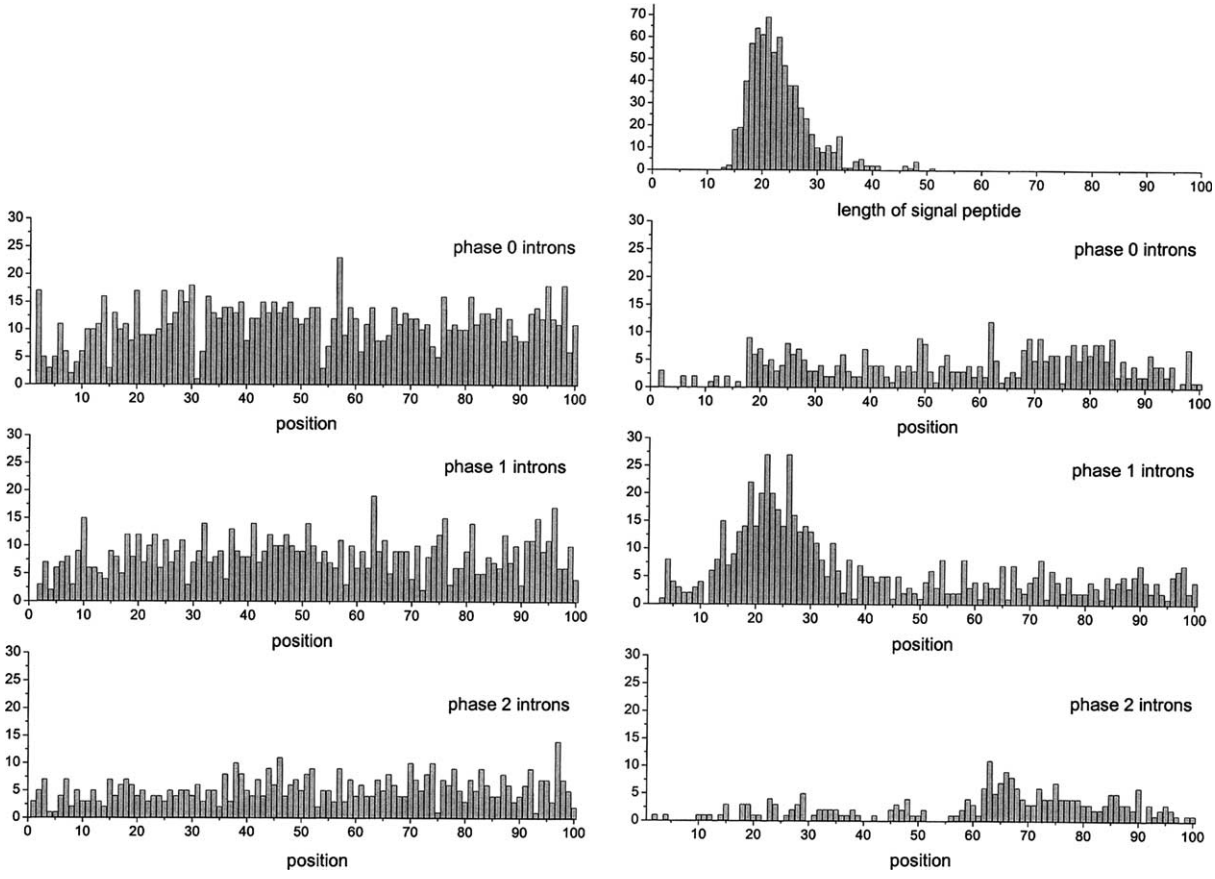

Fig. 1. Frequency distribution of phase 0, phase 1 and phase 2 introns along the first 100 amino acid residues of the translated regions of 2208 human genes. Left panel: genes of 1497 human proteins lacking N-terminal signal peptides. Right panel: genes of 711 human proteins possessing N-terminal secretory signal peptides. The top row in the right panel shows the length distribution of signal peptides. Note that in panel B a peak of excess phase 1 introns coincides with the C-terminal boundary of signal peptide-domains.

there is a peak of excess phase 1 introns in the region corresponding to the N-terminal 15–25 amino acids, whereas the region corresponding to residues 1–15 is significantly depleted in phase 0 introns and phase 2 introns (Fig. 1B). It is very likely that both types of deviations reflect the presence of the signal peptides: phase 0 and phase 2 introns are depleted in the N-terminal hydrophobic regions of the signal peptides, whereas the phase 1 intron peak coincides with the signal peptide cleavage sites (cf. top row of Fig. 1B).

The most plausible explanation for the influence of signal peptide domains on intron phase distributions is that some features of signal peptides favor intron insertion in phase 1 in the vicinity of the cleavage site, whereas other features disfavor insertion of introns in the N-terminal hydrophobic part.

As discussed in Section 1, introns are most likely to be inserted into genes at so called "proto-splice sites": sequences similar to the conserved coding sequences that usually flank introns [16]. Recent statistical analysis of a large number of exon–intron junction sequences has revealed that the proto-splice site is practically restricted to AG/G surrounding the intron, the information content being highest for the positions adjacent to the introns [24]. In translated regions, the base preferences of proto-splice sites (AG/G) are also manifested in some amino acid preferences at exon–intron junctions [23]. As a consequence, phase 2 introns most frequently split the arginine codon AGG, whereas phase 1 introns most frequently split the four GGN codons encoding glycine [23].

An analysis of the cleavage site matrices of eukaryotic signal peptides indicates that glycine residues are significantly enriched in positions −1, −3, −4 and −5 [25] (http://www.cbs.dtu.dk/services/SignalP/sp_matrices.html). It seems thus likely that the peak of phase 1 introns in the vicinity of signal peptide cleavage sites is a consequence of the abundance of glycines in this region, which present phase 1 proto-splice sites and this favors the insertion of introns in phase 1.

Conversely, the predominance of leucine in the hydrophobic N-terminal part of eukaryotic signal peptides from positions −13 to −6 [25] (http://www.cbs.dtu.dk/services/SignalP/sp_matrices.html), may explain the fact that there are significantly fewer phase 0 and phase 2 introns in this region than expected by chance (cf. Fig. 1B). Since none of the six codons of leucine (CTN, TTR) satisfy the proto-splice-site consensus sequence AG/G in either phase 0 or phase 2, the leucine-rich region is a very poor target for insertion of perfect spliceosomal introns in phase 0 or phase 2.

The fact that insertion of phase 1 introns at the boundary of signal peptide domains has been facilitated by the abundance of proto-splice sites in these regions has some important consequences for gene prediction. In the majority of cases, the signal peptide domains of secreted proteins or transmembrane proteins are encoded by separate exons (cf. Table 1.) making their detection one of the major problems of gene prediction protocols.

References

[1] Palmer, J.D. and Logsdon Jr., J.M. (1991) The recent origins of introns. Curr. Opin. Genet. Dev. 1, 470–477.

[2] Logsdon Jr., J.M. (1998) The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. 8, 637–648.

[3] Lynch, M. and Richardson, A.O. (2002) The evolution of spliceosomal introns. Curr. Opin. Genet. Dev. 12, 701–710.

[4] Stiller, J.W., Duffield, E.C. and Hall, B.D. (1998) Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II. Proc. Natl. Acad. Sci. USA 95, 11769–11774.

[5] Fast, N.M. and Doolittle, W.F. (1999) Trichomonas vaginalis possesses a gene encoding the essential spliceosomal component, PRP8. Mol. Biochem. Parasitol. 99, 275–278.

[6] Archibald, J.M., O'Kelly, C.J. and Doolittle, W.F. (2002) The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. Mol. Biol. Evol. 19, 422–431.

[7] Nixon, J.E., Wang, A., Morrison, H.G., McArthur, A.G., Sogin, M.L., Loftus, B.J. and Samuelson, J. (2002) A spliceosomal intron in Giardia lamblia. Proc. Natl. Acad. Sci. USA 99, 3701–3705.

[8] Logsdon Jr., J.M., Stoltzfus, A. and Doolittle, W.F. (1998) Molecular evolution: recent cases of spliceosomal intron gain? Curr. Biol. 8, R560–R563.

[9] O'Neill, R.J.W., Brennan, F.E., Delbridge, M.L., Crozier, R.H. and Graves, J.A.M. (1998) De novo insertion of an intron into the mammalian sex determining gene, SRY. Proc. Natl. Acad. Sci. USA 95, 1653–1657.

[10] Robertson, H.M. (2000) The large srh family of chemoreceptor genes in *Caenorhabditis nematodes* reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. 10, 192–203.

[11] Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G. and Koonin, E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr. Biol. 13, 1512–1517.

[12] Sharp, P.A. (1985) On the origin of RNA splicing and introns. Cell 42, 397–400.

[13] Patthy, L. (1995) Protein evolution by exon-shuffling. Molecular Biology Intelligence Unit. R.G. Landes Company, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.

[14] Bonen, L. and Vogel, J. (2001) The ins and outs of group II introns. Trends Genet. 17, 322–331.

[15] Bhattacharya, D., Lutzoni, F., Reeb, V., Simon, D., Nason, J. and Fernandez, F. (2000) Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes. Mol. Biol. Evol. 17, 1971–1984.

[16] Dibb, N.J. and Newman, A.J. (1989) Evidence that introns arose at proto-splice sites. EMBO J. 8, 2015–2021.

[17] Stoltzfus, A., Logsdon Jr., J.M., Palmer, J.D. and Doolittle, W.F. (1997) Intron 'sliding' and the diversity of intron positions. Proc. Natl. Acad. Sci. USA 94, 10739–10744.

[18] Coghlan, A. and Wolfe, K.H. (2004) Origins of recently gained introns in *Caenorhabditis*. Proc. Natl. Acad. Sci. USA 101, 11362–11367.

[19] Qiu, W.G., Schisler, N. and Stoltzfus, A. (2004) The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol. Biol. Evol. 21, 1252–1263.

[20] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucl. Acids Res. 31, 365–370.

[21] Saxonov, S., Daizadeh, I., Fedorov, A. and Gilbert, W. (2000) EID: the Exon–Intron Database – an exhaustive database of protein-coding intron-containing genes. Nucl. Acids Res. 28, 185–190.

[22] Long, M., Rosenberg, C. and Gilbert, W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA 92, 12495–12499.

[23] Tomita, M., Shimizu, N. and Brutlag, D.L. (1996) Introns and reading frames: correlation between splicing sites and their codon positions. Mol. Biol. Evol. 13, 1219–1223.

[24] Long, M., de Souza, S.J., Rosenberg, C. and Gilbert, W. (1998) Relationship between proto-splice sites and intron phases: evidence from dicodon analysis. Proc. Natl. Acad. Sci. USA 95, 219–223.

[25] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1–6.